A dataset for automatic detection of places in (early) modern French texts

Simon Gabay, Pedro Javier Ortiz Suárez

A dataset for automatic detection of places in (early) modern French texts

Simon Gabay[1] and Pedro Ortiz Suárez[2]

[1]University of Geneva (Switzerland)
[2]INRIA/Sorbonne Université (France)

With the apparition of distant reading (Moretti 2013), graphs, trees but also maps (cf. fig. 1) are becoming slowly a new way to "read" (Moretti 2005). But if it is easy to manually list places in a text, it is obviously too complicated to do the same on huge corpora. In order to accelerate the extraction of the information and simplify the production of such visualisations, it is therefore crucial to detect computationally place names.

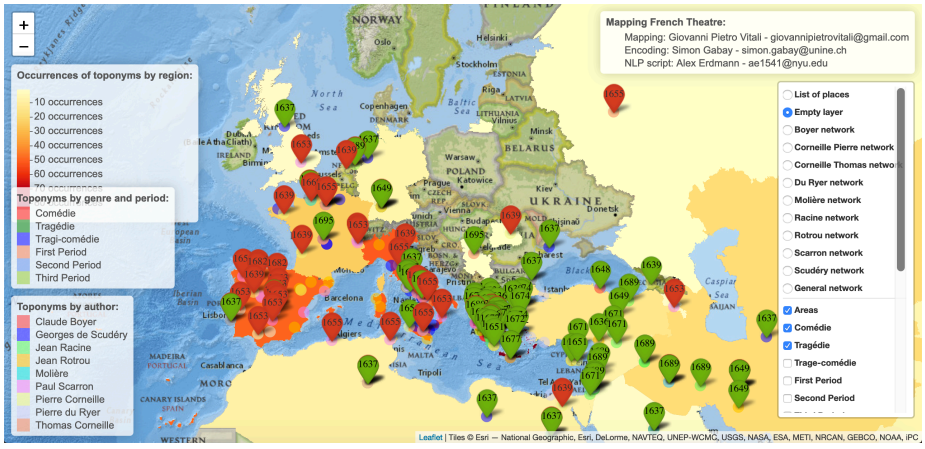

Figure 1: Place names mentioned in 17th c. place names (Gabay and Vitali 2019)

Sophisticated tools exist for contemporary French (Ortiz Suárez et al. 2020), but their results are heavily affected by spelling variation in historical documents, for which no tool exist (Gabay and Vitali 2019). Indeed, no matter how close they are, for a computer, *geneve* is not the same as *Genève* (Kudo and Richardson 2018) (cf. exp. 1). A short-term answer to the problem is the alignment of old forms with contemporary ones (Kogkitsidou and Gambette 2020), but normalisation techniques are far from being reliable enough despite recent improvements in the field (Gabay and Barrault 2020).

il établit des évêques en *Dannemarck*	il établit des évêques en *Danemark*
je te sais à *Pierre Sise*	je te sais à *Pierre Scize*

Example 1: Example of linguistic normalisation

Following our previous work at the crossroads of natural language processing and 17th c. French literature (Gabay, Bartz, and Deguin 2020; Gabay, Camps, and Clérice 2020), we have decided to create a dataset for (early) modern French (16th-18th c.) and make it available to specialists of machine learning to train new efficient models for Named Entity Recognition and Classification (NERC), but also Named Entity Linking (EL) (Frontini, Brando, and Ganascia 2015). To do so, a completely revised version of the *Presto* gold corpus (Gabay, Clérice, et al. 2020) of c. 4,900,000 lines is currently being annotated (cf. Appendix) using fine-grained entity types (following the Quaero guidelines, cf. Rosset, Grouin, and Zweigenbaum 2011) augmented, when possible, with wikidata IDs (Müller-Birn et al. 2015). This should enable researcher not only to locate place names in texts, but also to retrieve geographic coordinates to easily create maps.

Being orders of magnitude bigger than previously annotated corpora in historical French (Ehrmann et al. 2020), contemporary French (Sagot, Richard, and Stern 2012) and even English (Pradhan et al. 2012; Tjong Kim Sang and De Meulder 2003); this revised version of the *Presto* gold corpus could prove to be an useful tool for researchers in NLP (Natural Language Processing) given already very promising results even when used to train simplistic NER models (Lample et al. 2016).

Finally, if training NERC and EL models is a task for computer scientists, the creation of corpora and their annotation should be done in collaboration by specialists of the humanities. Preparing datasets being extremely time consuming, it is however important to avoid competing practices among researchers and projects, and therefore to engage the debate at the broader level to determine common annotation principles.

Data

Data are available at the following address: `github.com/e-ditiones/LEM17`.

Acknowledgments

Un grand merci à Giovanni Pietro Vitali (UVSQ) pour ses conseils en cartographie numérique.

Appendix

Grégoire	Grégoire	Np	B-pers	B-pers.ind	B-comp.name	O	Q133063
VII	7	Mc	I-pers	I-pers.ind	B-comp.qualifier	O	Q133063
était	être	Vuc	O	O	O	O	_
avec	avec	S	O	O	O	O	_
la	le	Da	O	O	O	O	_
comtesse	comte	Nc	B-pers	B-pers.ind	B-comp.title	O	Q464162
Matilde	Matilde	Np	I-pers	I-pers.ind	B-comp.name	O	Q464162
dans	dans	S	O	O	O	O	_
la	le	Da	O	O	O	O	_
ville	ville	Nc	B-loc	B-loc.adm.town	B-comp.kind	O	Q111144
de	de	S	I-loc	I-loc.adm.town	O	O	Q111144
Canosse	Canossa	Np	I-loc	I-loc.adm.town	B-comp.name	O	Q111144
,	,	Fw	O	O	O	O	_
l’	le	Da	O	O	O	O	_
ancien	ancien	Ag	B-loc	B-loc.adm.town	B-comp-qualifier	O	Q111144
Canusium	Canusium	Np	I-loc	I-loc.adm.town	B-comp.name	O	Q111144
,	,	Fw	O	O	O	O	_
sur	sur	S	O	O	O	O	_
l’	le	Da	O	O	O	O	_
Apennin	Apennin	Np	B-loc	B-loc.geo.phys	O	O	Q1285
près	près	Rg	O	O	O	O	_
de	de	S	O	O	O	O	_
Régio	Reggio	Np	B-loc	B-loc.adm.town	O	O	Q13360
,	,	Fw	O	O	O	O	_
forteresse	forteresse	Nc	O	O	O	O	_
qui	qui	Pr	O	O	O	O	_
passait	passer	Vvc	O	O	O	O	_
alors	alors	Rg	O	O	O	O	_
pour	pour	S	O	O	O	O	_
imprenable	imprenable	Ag	O	O	O	O	_
.	.	Fs	O	O	O	O	_

Les	le	Da	O	O	O	O	_
allemands	allemand	Nc	O	O	O	O	_
élurent	élire	Vvc	O	O	O	O	_
pour	pour	S	O	O	O	O	_
empereur	empereur	Nc	B-pers	B-pers.ind	B-comp.title	O	Q438435
Rodolphe	Rodolphe	Np	I-pers	I-pers.ind	B-comp.name	O	Q438435
duc	duc	Nc	I-pers	I-pers.ind	B-comp.title	O	Q438435
de	de	S	I-pers	I-pers.ind	I-comp.title	O	Q438435
Suabe	Souabe	Np	I-pers	I-pers.ind	I-comp.title	B-loc.adm.reg	Q438435

Example 2: NERC Fine-Grained annotation with EL

References

Ehrmann, M. et al. (2020). "Extended overview of CLEF HIPE 2020: named entity processing on historical newspapers". In: *CLEF 2020 Working Notes. Conference and Labs of the Evaluation Forum*. Vol. 2696. CONF. CEUR.

Frontini, F., C. Brando, and J.-G. Ganascia (June 2015). "Semantic Web Based Named Entity Linking for Digital Humanities and Heritage Texts". In: *First International Workshop Semantic Web for Scientific Heritage at the 12th ESWC 2015 Conference*. Ed. by A. Zucker et al. Arnaud Zucker and Isabelle Draelants and Catherine Faron Zucker and Alexandre Monnin. Portoro, Slovenia. URL: https://hal.archives-ouvertes.fr/hal-01203358.

Gabay, S. and L. Barrault (June 2020). "Traduction automatique pour la normalisation du français du XVII e siècle". In: *TALN 2020*. 27ème Conférence sur le Traitement Automatique des Langues Naturelles. ATALA. Nancy, France. URL: https://hal.archives-ouvertes.fr/hal-02596669.

Gabay, S., A. Bartz, and Y. Deguin (Oct. 2020). "CORPUS17: a philological corpus for 17th c. French". In: *Proceedings of the 2nd International Digital Tools & Uses Congress (DTUC '20)*. Hammamet, Tunisia. DOI: 10.1145/3423603.3424002. URL: https://hal.archives-ouvertes.fr/hal-03041871.

Gabay, S., J.-B. Camps, and T. Clérice (May 2020). *Manuel d'annotation linguistique pour le français moderne (XVIe -XVIIIe siècles)*. Manuel d'annotation en vue de la création de modèle de lemmatisation et d'annotation morpho-sytaxique et morphologique du français des XVI-XVIIIème s. URL: https://hal.archives-ouvertes.fr/hal-02571190.

Gabay, S., T. Clérice, et al. (Oct. 2020). "Standardizing linguistic data: method and tools for annotating (pre-orthographic) French". In: *Proceedings of the 2nd International Digital Tools & Uses Congress (DTUC '20)*. Hammamet, Tunisia. DOI: 10.1145/3423603.3423996. URL: https://hal.archives-ouvertes.fr/hal-03018381.

Gabay, S. and G. P. Vitali (Nov. 2019). "A Theatre of Places: Mapping 17 th French Theatre". In: *GIR'19 - 13th Workshop on Geographic Information Retrieval*. Proceedings of the 13th Workshop on Geographic Information Retrieval. Lyon, France: ACM. DOI: 10.1145/3371140.3371146. URL: https://hal.archives-ouvertes.fr/hal-02388411.

Kogkitsidou, E. and P. Gambette (Nov. 2020). "Normalisation of 16th and 17th century texts in French and geographical named entity recognition". In: *ACM SIGSPATIAL GeoHumanities'20*. Proceedings of the 4th ACM SIGSPATIAL Workshop on Geospatial Humanities. ACM. Seattle (virtual), United States, pp. 28–34. DOI: 10.1145/3423337.3429437. URL: https://hal-upec-upem.archives-ouvertes.fr/hal-02955867.

Kudo, T. and J. Richardson (Nov. 2018). "SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Brussels, Belgium: Association for Computational Linguistics, pp. 66–71. DOI: 10.18653/v1/D18-2012. URL: https://www.aclweb.org/anthology/D18-2012.

Lample, G. et al. (June 2016). "Neural Architectures for Named Entity Recognition". In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics, pp. 260–270. DOI: 10.18653/v1/N16-1030. URL: https://www.aclweb.org/anthology/N16-1030.

Moretti, F. (2005). *Graphs, maps, trees : abstract models for a literary history*. Londres: Verso. ISBN: 978-1-84467-185-4.

— (2013). *Distant reading*. London: Verso. ISBN: 978-1-78168-084-1.

Müller-Birn, C. et al. (2015). "Peer-production system or collaborative ontology engineering effort | Proceedings of the 11th International Symposium on Open Collaboration". In: *OpenSym '15: Proceedings of the 11th International Symposium on Open Collaboration*, pp. 1–10. URL: https://dl.acm.org/doi/10.1145/2788993.2789836 (visited on 02/08/2021).

Ortiz Suárez, P. J. et al. (May 2020). "Establishing a New State-of-the-Art for French Named Entity Recognition". English. In: *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, pp. 4631–4638. ISBN: 979-10-95546-34-4. URL: https://www.aclweb.org/anthology/2020.lrec-1.569.

Pradhan, S. et al. (July 2012). "CoNLL-2012 Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes". In: *Joint Conference on EMNLP and CoNLL - Shared Task*. Jeju Island, Korea: Association for Computational Linguistics, pp. 1–40. URL: https://www.aclweb.org/anthology/W12-4501.

Rosset, S., C. Grouin, and P. Zweigenbaum (2011). *Entités nommées structurées : guide dannotation Quaero*. URL: http://www.quaero.org/media/files/bibliographie/quaero-guide-annotation-2011.pdf.

Sagot, B., M. Richard, and R. Stern (June 2012). "Annotation référentielle du Corpus Arboré de Paris 7 en entités nommées (Referential named entity annotation of the Paris 7 French TreeBank) [in French]". In: *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 2: TALN*. Grenoble, France: ATALA/AFCP, pp. 535–542. URL: https://www.aclweb.org/anthology/F12-2050.

Tjong Kim Sang, E. F. and F. De Meulder (2003). "Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition". In: *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pp. 142–147. URL: https://www.aclweb.org/anthology/W03-0419.